Middlesex University London

Middlesex University Research Repository

An open access repository of

Middlesex University research

http://eprints.mdx.ac.uk

Juddoo, Suraj and George, Carlisle ORCID logoORCID: https://orcid.org/0000-0002-8600-6264 (2018) Discovering the most important data quality dimensions in health big data using latent semantic analysis. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). In: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 06-07 Aug 2018, Durban, South Africa. ISBN 9781538630600. [Conference or Workshop Item] (doi:10.1109/ICABCD.2018.8465129)

Final accepted version (with author's formatting)

This version is available at: https://eprints.mdx.ac.uk/25560/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this work are retained by the author and/or other copyright owners unless otherwise stated. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge.

Works, including theses and research projects, may not be reproduced in any format or medium, or extensive quotations taken from them, or their content changed in any way, without first obtaining permission in writing from the copyright holder(s). They may not be sold or exploited commercially in any format or medium without the prior written permission of the copyright holder(s).

Full bibliographic details must be given when referring to, or quoting from full items including the author's name, the title of the work, publication details where relevant (place, publisher, date), pagination, and for theses or dissertations the awarding institution, the degree type awarded, and the date of the award.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:

eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.

See also repository copyright: re-use policy: http://eprints.mdx.ac.uk/policies.html#copy

Presented at the *IEEE International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) Durban, South Africa, 6-7 August 2018*

Discovering the Most Important Data Quality Dimensions in Health Big Data Using Latent Semantic Analysis

Suraj Juddoo
School of Science and Technology
Middlesex University Mauritius
Cascavelle, Mauritius
Email: s.juddoo@mdx.ac.mu

Dr Carlisle George
School of Science and Technology
Middlesex University,
London, UK
Email: c.george@mdx.ac.uk

***Abstract*-**Big Data quality is a field which is emerging. Many authors nowadays agree that data quality is still very relevant, even for Big Data uses. However, there is a lack of frameworks or guidelines focusing on how to carry out big data quality initiatives. The starting point of any data quality work is to determine the properties of data quality, termed 'data quality dimensions' (DQDs). Even these dimensions lack precise rigour in terms of definition in existing literature. This current research aims to contribute towards identifying the most important DQDs for big data in the health industry. It is a continuation of previous work, which, using relevant literature, identified five DQDs (*accuracy, completeness, consistency, reliability and timeliness*) as being the most important DQDs in health datasets. The previous work used a human judgement based research method known as an inner hermeneutic cycle (IHC). To remove the potential bias coming from the human judgement aspect, this research study used the same set of literature but applied a statistical research method (used to extract knowledge from a set of documents) known as latent semantic analysis (LSA). Use of LSA concluded that *accuracy* and *completeness* were the only similar DQDs classed as the most important in health Big Data for both IHC and LSA.

Keywords-Big Data, data quality, health data, data quality dimensions, latent semantic analysis.

I. INTRODUCTION

Big Data originates from situations when traditional relational database technologies can no longer meet the requirements of users due to the amount, varying types and incoming streams of data [1]. However, other authors are adding new properties amongst which 'Value' and 'Veracity' are the most prominent. Veracity concerns the rising issue of certainty or quality involved with using data. Data quality (DQ) is generally explained as data 'fit for use'. This broad definition conveys the notion that data is used for certain objectives, and quality data would be data which would be adequate enough to allow the users of data to meet their objectives. In the realm of Big Data, research about quality is still at an infancy stage, offering opportunities to probe deeper into understand implications and impacts.

The healthcare sector is an industry that collects a large amount of data. Health data is essential for the proper delivery of health services and can be in different formats: electronic health records, administrative data, claims data, disease registries, health surveys and clinical trials data. Apart from Google Flu and Ebola forecasts, there is a growing number (but lesser known) ways in which Big Data are being used in the health industry. Examples include use of Big Data to reduce patient readmissions or to more accurately identify diseases such as cancer.

The first step in any data quality activity is to express the properties of data quality; those properties are termed data quality dimensions (DQDs) [2][17][8][5]. Examples of very frequently cited DQDs involved with Big Data are: consistency, accuracy, completeness, timeliness [9]. The principal goal of this paper is to discuss, analyse and recommend DQDs suitable in the context of very large health datasets. Investigations into DQDs is still a currently very important consideration, especially in the context of Big Data. A survey of literature indicates that there is a lack of standard framework of what constitutes the most important DQDs [12]. This lack of clarity for DQDs is even more noticeable within the context of Big Data, as this field of research is new and there exists very few Big Data research dedicated towards data quality.

A past unpublished research study conducted by the current authors, used an inner hermeneutic cycle (IHC), a very accepted form of integrative literature review [16] to identify the most important DQDs for big datasets in the health industry. The study concluded that *accuracy*, *completeness*, *consistency*, *reliabilit*y and *timeliness* are the most important DQDs. The aim of this current research study was to use a different research method with less bias from author interpretation, to identify the most important DQDs in the same literature corpus as the previous research study. The different research method used was a statistical technique specialized in identifying word-word, word-document and document-document relationships known as latent semantic analysis (LSA). The implementation of LSA in this current work was carried out using python 2.7 libraries. A final aim of this work was to compare the results between use of the IHC and LSA research methods.

II. LITERATURE REVIEW

I. Big Data quality dimensions

The high volume and velocity properties of Big Data can result in poor data quality. Furthermore, due to data coming from multiple sources, data integrity and consistency concerns increase. Many proposals have been made to improve data quality, for example, through machine learning approaches to avoid depending upon users and their inherent weaknesses [13]. In the early days of Big Data, some authors argued that the importance of improving data quality for Big Data might not be high, since the amount of incorrect data was deemed to be negligible and would not affect the final outcome of data analysis [15]. Currently, there is a much wider acceptance of the importance of data quality for Big Data in general. A better understanding of data quality dimensions that is more relevant for Big Data will support research and industry in building more appropriate data quality tools.

One of the rare specialised research studies on DQDs of Big Data posits that the main data quality dimension to be considered for Big Data is consistency [4], which is defined as, *the capability of information systems to ensure uniformity of datasets when data are being transferred across networks and systems*. The main hypothesis is that the business value of a dataset can be estimated only in its context of use and therefore, the importance of data differs according to different data use. Consequently, asserting quality properties for data would change according to the different purposes of data use. The authors [4] further subdivided consistency into three subsequent parts, as discussed below and seen in Table 1. Additionally, they connected many of the traditional data quality dimensions with the three consistency subdomains as follows:

Contextual consistency refers to how far big datasets are used within same domain of interest independently of data format, size and velocity of production of data. For the current research, the domain of interest is health data. Relevancy, credibility, ease of understanding, accuracy and confidentiality are key DQ dimensions for this type of consistency.

Temporal consistency conveys the idea that data needs to be understood in a consistent time slot, such that the same data might not be comparable if they are from another time slot. Time concurrency, availability and currency are deemed to be essential for temporal consistency.

Operational consistency brings in the operational influence of technology upon the production and use of data. There are many sources of data in Big Data scenarios, hence operational consistency is crucial for ensuring veracity of Big Data. Availability, portability, precision, completeness and traceability are considered the main connected dimensions for this subtype.

Reference [4] mapped how the 3v's of big data affect the 3Cs of data quality as seen in Table 1:

Table 1: Matrix of 3Cs relative to the 3Vs [4]

	Velocity	**Volume**	**Variety**
Contextual	Consistency, Credibility, Confidentialiy	Completeness, Credibility	Accuracy, Consistency, understandability
Temporal	Consistency, Credibility, Currentness, Availability		Compliance
Operational	Completenes, Accessibility, Efficiency, Traceability, Availability, Recoverability	CompletenessA ccessibility, Efficiency, Availability, Recoverability	Accuracy, Compliance, Accessibility, Efficiency, Traceability, Availability, Recoverability, Precision

Other research studies have focused on determining DQDs for Big Data using the inner hermeneutic cycle such as work by [2]. The context of the study by [2] differs from the current work in this paper since [2] used three main Big Data 'coordinates' namely: data types, sources and application domains. They focused on maps, semi-structured texts, linked open data, sensor & sensor networks and official statistics. Correlations between DQDs and the Big Data coordinates were reported as: *accuracy* for maps, *completeness* of official statistics, *readability* for semi-structured data, *accessibility and trust* for linked open data and *consistency* for sensor and sensor networks. The authors in [2] performed an inner hermeneutic cycle comprising of an initial corpus of 1600 papers, related tables and notes. Keywords as part of the titles and abstracts having a minimum thread of 100 citations for the period of 2005 to 2014 were used as the criteria for sorting. A summary of the literature review results were used to devise their theoretical conceptual framework which detailed Big Data quality dimensions clustered by the above cited application areas ranging from maps to official statistics.

Other authors developed a framework for both data-driven and processed driven data quality aspects of big data [4]. They evaluated quality initiatives into two distinct components, firstly the data quality intrinsically and secondly, the processes of handling data. As part of the data quality evaluation, they also focused on pre-processing activities such as data cleansing. They experimented with health big datasets amongst others and specified only three main DQDs as part of their data-driven aspects namely: accuracy, completeness and consistency.

II. Latent Semantic analysis (LSA)

LSA is a statistical method for estimating the meaning of terms based on linear combinations of underlying concepts. It has been applied in a variety of fields ranging from operations research management, library indexing improvement, and search engine query performance optimisation to chatters' perceptions on social networks [10]. The fact is that wherever meaning or importance of terms need to be extracted from a set of text data, LSA is a technique worth considering. LSA is a technique created decades ago, in late 1980's/ early 1990's. It had been primarily applied extensively in the field of search engine performance optimisation, with the aim of helping users to more precisely find appropriate search results based on specific search queries. It is still a very relevant method due to the 'variety' property of Big Data, which is the production of a huge amount of structured and unstructured data which are logically related. Decision makers want to have the ability to work with all these data together, but the semantics or choice of terms used might be different according to different authors of documents and therefore, there should be ways to create inductive relationships between terms and documents.

LSA has been applied in operations management with the purpose of uncovering the intellectual structure of the domain area [10]. The authors in [10] used abstracts coming from highly rated research journals specialised in the area of operations management (OM). Given the interdisciplinary nature of some journals, care was taken to select only abstracts related to OM. 3207 abstracts obtained from the EBSCO library from the time period 1980-2012 were used for LSA processing. The primary goal was to identify the core research topics that scholars in the field of OM focused upon during the last three decades approximatively. A vocabulary of 1078 stemmed terms were used to develop the term matrix over the 3207 documents. Subsequently, term frequency–inverse document frequency (TD-IDF) transformations were carried upon the term matrix document, followed by the application of single value decomposition (SVD), which is mathematical decomposition technique very similar to factor analysis and mainly recommended for text analysis. The authors made a choice of selecting the number and choice of factors to consider for the LSA analysis based on results of trial and error parameter setting.

Another example of an area where LSA has been applied involved the identification of emotions from comment texts when images are uploaded on social networks [18]. LSA was the chosen method for this research study as it is described as a light weight method, that is, a method which consumes relatively little computing and memory resources compared to sentiment analysis techniques. As LSA is a statistical method which does not require any training model, it was chosen compared to other machine learning based alternatives. The authors in [18] performed the typical steps involved with LSA application in terms of text pre-processing, term document matrix formulation and finally the application of SVD logic for dimensionality reduction. Six main terms were factored in from a corpus of 27 texts extracted from "psychpage.com". The average accuracy of emotion association was 81.48%.

III. RESEARCH METHODOLOGY

LSA has been used in previous research studies to determine the meaning of words and passages of large text corpora [11]. LSA applies an SVD algorithm to reduce the dimensional representation of a text matrix of words to documents, ultimately resulting in the importance of a given word for a corpus of documents. With SVD, the importance of a word in a context might be greater compared to the count of the word in the same context, because SVD would forecast the importance of the particular word in a projected infinite amount of articles. Alternatively, even if a word appears frequently in a particular research paper, the application of SVD might result in a low importance if the factor analysis algorithm predicts that this frequent occurrence is only for this specific research paper and might not hold subsequent importance for the whole research domain area. LSA determines the similarity of the meaning of words and set of words inside a large corpus of text which is traditionally an activity based on human interpretation. Empirical experience supports the use of LSA as an alternative for human interpretation: an example is the application of LSA to estimate the quality and quantity of knowledge in essays where LSA was highly comparable to human interpretation [11].

Hence, with the application of LSA, this research study aimed to determine the order of importance of thirty-eight (38) data quality dimensions uncovered via the application of IHC in the unpublished study (previously mentioned), using the same set of literature. LSA evaluated the similarity occurrence of those individual 38 data quality dimensions found in the previous research study by the application of a cosine similarity measure.

IV. LSA IMPLEMENTATION

The application of LSA for this research study made use of a corpus of 34 research abstracts which were selected from various articles focusing on data quality, Big Data and health informatics. The same corpus was used for the previous unpublished research study (mentioned above) with the common aim of identifying the most important DQDs in the area of very large health datasets. However, due to issues such as some of the articles not having an abstract section and the limits of memory processing capacity of the 'MatrixSimilarity' method of the Gensim package, only 34 abstracts were used. The decision to use only abstracts for detecting relationships relative to terms within documents was adopted from an extremely authoritative work in the field of LSA application [6]. Additionally, the LSA algorithms performs forecasting of the term to document similarity in a much more complex way compared to simple comparisons such that the essential writings of a research document congested within an abstract should help determine the similarity importance.

Step 1: A corpus of raw text was obtained. The documents containing previous literature in this particular research area consisted of 43 documents in pdf format but only 34 contained an acceptable abstract section, hence this research study had to limit itself to using 34 documents. The first step was to apply an algorithm to convert .pdf documents into text format (.txt). This was undertaken using the "pdfminer" class of python. However, during this process, figures and charts were not converted into text. Furthermore, characters such as '=,<,>' generated a compiler error with the genism package. Those characters were removed from the documents inserted as input to the LSA algorithm without any potential consequences for the LSA results since they did not show any link to DQDs. Also, some pdf articles were produced as image based articles, and therefore could not be directly converted to text format. Therefore, a non-pdf equivalent was obtained from research databases and the text equivalent of the abstract was extracted. The total size of the 34 documents was processed using a laptop with 8Gb of RAM. The algorithm was able to identify the documents using a zero based indexing system. Table 2 below provides a list of index numbers mapped to research article titles:

Table 2: Mapping of index numbers to research article

Index	Research title
0	A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical research
1	Completeness and accuracy of data transfer of routine maternal health services data in the greater Accra region
2	A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments
3	Valid comparisons and decisions based on clinical registers and population based cohort studies: assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database
4	Accuracy of injury coding under ICD-9 for New Zealand public

	hospital discharges
5	Big Data Quality: A Quality Dimensions Evaluation
6	An Hybrid Approach to Quality Evaluation Across Big Data Value Chain
7	FROM DATA QUALITY TO BIG DATA QUALITY
8	Challenges in data quality: the influence of data quality assessments on data availability and completeness in a VMMC programme in Zimbabwe
9	Classifying, measuring and improving the quality of data in trauma registries: A review of the literature
10	Creating a General (Family) Practice Epidemiological Database in Ireland - Data Quality Issue Management
11	Data Challenges in Disease Response: The 2014 Ebola Outbreak and Beyond
12	DATA MINING CONSULTING IMPROVE DATA QUALITY
13	Data Quality: A Survey of Data Quality Dimensions
14	Data Quality by Contract – Towards an Architectural View for Data Quality in Health Information Systems
15	Data Quality Problems When Integrating Genomic Information
16	Data representation factors and dimensions from the quality function deployment (QFD) perspective
17	Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research
18	A Methodology to Evaluate Important Dimensions of Information Quality in Systems
19	The influence of calibration method and eye physiology on eyetracking data quality
20	Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies
21	Contrasting the Dimensions of Information Quality in their Effects on Healthcare Quality in Hospitals
22	Efficient quality-driven source selection from massive data sources
23	Measuring the quality of patient data with particular reference to data accuracy
24	Open data quality measurement framework: Definition and application to Open Government Data
25	Does use of computer technology for perinatal data collection influence data quality?
26	Identifying Relationships of Information Quality Dimensions
27	The Challenges of Data Quality and Data Quality Assessment in the Big Data Era
28	The Effects and Interactions of Data Quality and Problem Complexity on Classification
29	The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases
30	Transparent Reporting of Data Quality in Distributed Data Networks
31	A Pilot Ontology for a Large, Diverse Set of National Health Service Healthcare Quality Indicators
32	Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work
33	Discovering Dependencies among Data Quality Dimensions: A Validation of Instrument

Step 2: The text document was pre-processed for analysis. Firstly, this involved eliminating unnecessary characters such as page numbers, symbols and white spaces. Secondly, a stop list of words that needed to be excluded from analysis had to be devised. The full list used is ***"for a of the and to namely higher in on at data their ours yours her his and from other are with such but require is care We we They these using over can that towards within between known be users 0 1 2 3 4 5 6 7 8 9 & first were was also %"***. The algorithm developed was case sensitive, therefore, some words had to be inserted in both lower and upper case. Most of the words were common trivial English words which did not have a major difference in the semantic structure of the documents. Without the stop list, the words on it would not have been discarded by the LSA algorithm which basically creates an index based on more than one occurrence inside the whole corpus of 34 documents/abstracts. Thus, their specification as terms to be discarded during the first cycle of processing was essential. This stop list could have included many other words which according to human judgement would not have had any kind of major impact on the semantic structure of the documents relative to data quality dimensions. However, because LSA takes into account the semantic structure in a much more complex way compared to human judgement (i.e. not just comparing exact query terms but having some global evaluation of the terms within a corpus of documents) the decision was taken to only specify extremely common English terms and special characters. Subsequently, word stemming was performed using a pre-built word stemming python class known as NLTK.

Step 3: A term document matrix was created. This is a row-column tabular representation of counts of terms per document. The columns represented the document, which referred to research papers in the area of DQDs for big health datasets. The rows represented the terms, which were the 38 data quality dimensions. Python 2.7 was used, along with the *Gensim* package. A dictionary was built from the 34 documents which resulted in 706 unique tokens, that referred to terms extracted from the documents. The Gensim package contained classes allowing the TF-IDF transformation and the dimensionality reduction for the SVD. After application of the inverse document frequency (IDF) weights, a 2120 matrix of non-sparse elements was created.

Step 4: The dimensionality of the term matrix document was reduced. The SVD class of LSA was applied to the term matrix document in order to relate the importance of terms per document. The SVD application was carried out in two phases. During the first phase, a 706 by 110 action matrix was constructed and subsequently 'orthonormalised'. During the second phase, a dense SVD was carried out and created a 110 by 34 action matrix. Ten factors were kept by the LSA algorithm which resulted in the elimination of 59.7% of the tokens as per step 3 above. An example of the output per document was as follows for document 0: topic #0(1.607): -0.385*"big" + -0.142*"information" + -0.135*"assessment" + -0.135*"dimensions" + -0.108*"paper" + -0.102*"health" + -0.102*"have" + -0.099*"framework" + -0.093*"research" + -0.088*"clinical".

Finally, a cosine similarity function was computed for the index created from the 34 documents with the use of 10 latent dimensions. For the search query term 'Completeness', the following result was produced:

[(8, 0.92762876), (1, 0.71302426), (3, 0.62934136), (2, 0.58216494), (22, 0.55686307), (29, 0.48969993), (4, 0.38626587), (18, 0.36283055), (11, 0.35054082), (30, 0.34343559), (20, 0.3291077), (23, 0.32587472), (9, 0.27751592), (6, 0.23394442), (5, 0.19664758), (15, 0.17808378), (25, 0.15371184), (10, 0.11480794), (17, 0.068790123), (26, 0.054444589), (28, 0.046295159), (0, 0.029957294), (19, 0.018134167), (7, 0.010269118), (13, -0.00048203743), (27, -0.014438681), (31, -0.079807945), (21, -0.094874345), (12, -0.12727518), (14, -0.1801782), (16, -0.18578127), (24, -0.20913719), (32, -0.24391271)]

V. ALGORITHM CREATED

The pseudocode behind the LSA implementation was as follows:

Documents = text abstracts of 34 research articles

Specify the stop list of words

Retrieve all words one by one from the documents

If a word appears more than once and not part of stop list, then add it as a token.

Create a dictionary with all individual tokens

Apply TF-IDF upon all the tokens

Apply Lsimodel method upon the term document matrix

Specify the matching/search query term

Apply MatrixSimilarity method to generate an index

Sort the index and display

VI. ANALYSIS AND FINDINGS

The algorithm above was applied for 38 DQDs, where individual DQDs were used as search query terms. The resulting values represented the cosine similarity of terms (DQDs) to documents (research abstracts). The cosine similarity represented the importance of a term per document, without taking into account individual word counts of each document. Out of the 38 dimensions used as query documents, 30 did not return any level of similarity or difference with the document corpus being examined. Hence, for all documents making up the corpus, the cosine similarity indicated 0. Those DQDs were hence discarded from further analysis.

A. Comparison between IHC and LSA results

The first major difference between the LSA and the IHC results (from the previous unpublished study) was that ***only 8 out of possible 38*** DQDs showed some similarity with the 34 documents forming the corpus. This may be explained by two factors: (1) the difference between the semantic interpretation of text between a human reader and LSA and; (2) the fact that only abstracts were analysed by the LSA algorithm and not the whole text. However, when considering the fact that the main reason for applying LSA was to determine which terms were important within a corpus of documents, the choice of only analysing the abstract section of an article was justified since the abstract is supposed to embody the main ideas present in any article. Therefore, by rejecting the second explanation given above it is possible to conclude that some DQD such as timeliness, which had a score of 11 for the IHC and 0 for LSA, did not have sufficient similarity when considering the whole corpus of abstracts merged together. This is because LSA does not work as a full text query for words but rather considers the importance of terms in a holistic interpretation of the corpus. This holistic interpretation may give some very surprising results such as: (1) some terms which appear a lot might not result in a high cosine similarity; (2) some terms which does not appear at all in some documents might show some similarity with those documents; (3) some documents which did contain certain terms display zero cosine similarity and; (4) high cosine similarity for some terms per specific document.

The fact that some DQDs did not reflect any match after the LSA application may be explained by number 3 above. In the previous IHC study the timeliness DQD was the fourth most important DQD. The IHC detected 8 research articles mentioning timeliness. 7 abstracts from these 8 articles were use in the LSA application. However, the cosine similarity matches were 0 for all research abstracts. This proves that with LSA, it is not individual word counts which were considered, but the semantic strength and relationship of terms within an overall corpus of documents.

The scatter plot chart shown in Figure 1 plots the 8 identified DQDs by the LSA from the 34 different documents represented by the x-axis. The y-axis denote the cosine similarity index within a range of -1 to 1, where the closer an index is to 1, the more important is the term for a particular corpus.

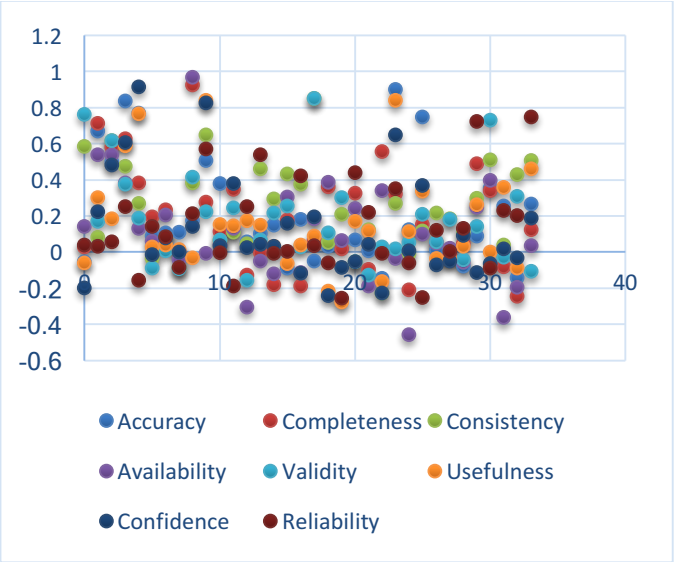

Figure 1. Scatter plot of cosine similarity values for DQDs per research article

Taking only a subsection of the cosine similarity which is greater than 0.8 as a benchmark for the most important DQDs with the cosine similarity index, we denote the following facts as shown in Table 3 below.

Table 3. Hierarchy of DQDs per cosine similarity

DDQs	Counts of cosine similarity > 0.8
Accuracy	2
Usefulness	2
Confidence	2
Availability	1
Validity	1
Completeness	1
Consistency	0
Reliability	0

As seen in Table 3, when using LSA, ***Accuracy, Usefulness and Confidence*** are the top 3 most important DQDs followed by ***Completeness, Availability and Validity.*** This hierarchy is quite different from the previous IHC results which concluded that the most important DQDs were: *accuracy, completeness, consistency, reliability and timeliness*. On the other hand,

accuracy is identified as one of the topmost DQD by both LSA and IHC.

Analysing the distribution of the cosine similarity indexes per range bands in Figure 1 gives another interesting insight. There are significantly many more similarity plots between the range of 0 to 0.4 compared to the range of 0.6 to 1. This means that even if some DQDs had been identified across the corpus of abstracts, their importance is largely ranked from low (0) to medium (0.5). One reason for this phenomena could be the fact that most of the 34 research articles discussed data quality, but do not necessarily focused on DQDs as their main locus of research. However, on comparison with the IHC results, the same trend could be discerned; with the IHC, there were 5 DQDs with a weighted count greater than 10 and most of the 33 other DQDS had very low weighted counts.

VII. CONCLUSION

This study focused on identifying the most important DQDs for health data sets and comparing the results with a previous study. The previous research study was based on human interpretation and applied an integrative literature review research method known as inner hermeneutic cycle (IHC). The IHC method has been widely used in various other research studies where knowledge needed to be extracted from existing literature. Since the IHC method is based on human interpretation, this may negatively impact on generalising the results. Hence, the current research study focused on applying a different statistical research method (called Latent Semantic Analysis) to the same body of literature

The statistical process of LSA and human interpretation using IHC produced different results. Whereas with IHC, 38 different DQDs with varying levels of importance were found, with LSA only 8 DQDs showed some connection with the overall corpus of literature, but again with varying levels of importance. With IHC, *accuracy, completeness, consistency, reliability and timeliness* were found to be the most important in ascending order. However, with LSA, *accuracy, usefulness, completeness, availability and validity* were found to be most important in ascending order. The two common most important DQDS found by the two different research methods were ***accuracy and completeness.*** Thus, the first principal conclusion based on comparing the results of using the two different research methods is that accuracy and completeness are the most important DQDs to consider in the context of big datasets for the health industry. Use of the two research methods also confirmed the fact that most of the other DQDs are not identified as 'very important' in this particular data and industry context.

This work forms part of a wider research project focused on optimising big data quality in the health industry. In the future experimental health datasets will be evaluated for their degree of accuracy and completeness, and machine learning techniques will be developed to classify accurate and complete data from inaccurate and incomplete data. Ultimately, a data repair algorithm would be developed to improve on the accuracy and completeness of the health big datasets used.

VIII. REFERENCES

[1] Aisling O'Driscoll, J. D. R. D. S. "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics,* 2013. pp. 774-781.

[2] Batini, C., Rula, A., Scannapieco, M. & Viscusi, G.. "FROM DATA QUALITY TO BIG DATA QUALITY," *Journal of Database Management,* Volume 1, pp. 60-82,2015

[3] BMJ. "Evidence based medicine: what it is and what it isn't," *BMJ,* 312(71), 1996

[4] Caballero, I., Serrano, M. & Piattinni, M. "A data quality in Use model for Big Data," ER workshops, pp. 65-74, 2014

[5] Cai, L. & Zhu, Y. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal,* 14(2), 2015

[6] Deerwester,S.,Dumais,S.,Furnas,G.,Landauer,T. & Harshman, R. *Indexing by Latent Semantic Analysis.* JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. 41(6):391-407, 1990

[7] Handler, D. J. "Small Data-Thinking Kills Big Data-Aspirations". http://www.wired.com/insights/2013/01/small-data-thinking-kills-big-data-aspirations/,2012.

[8] Huang, H., Stvilia, B. & Bass, H. "Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,* 2012.

[9] Juddoo,S. "Overview of Big Data quality challenges," *ICCCS IEEExplore conference proceedings*. doi: 10.1109/CCCS.2015.7374131,2015.

[10] Kulkarni, S.S.,Apte, U.M. & Evangelopoulos,N.E. *The use of Latent Semantic Analysis in Operations Management Research.* Journal of Decision Sciences Institute, Volume 45, No. 5, Oct 2014.

[11] Landauer, T. K., Foltz, P. W., & Laham, D. (1998).Introduction to Latent Semantic Analysis.*Discourse Processes,* **25**, 259-284.

[12] Pipino, L., Yang, L. & Wang, R. "Data Quality Assessment," *Communications of the ACM,*2002

[13] Saha, B. & Srivastava, D. "Data Quality: The other face of Big Data," AT&T Labs-Research, 2014

[14] Serhani, M.A.,Kassabi,H.T,Taleb,I. & Nujum,A. *An Hybrid Approach to Quality Evaluation Across Big Data Value Chain,* 2016 IEEE International Congress on Big Data, 2016

[15] Soares, S. "Big Data quality," *Big Data Governance: An emerging imperative.*MC Press, pp. 101-112, 2012

[16] Wang, R. & Strong, D. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems,* 12(4), pp. 5 - 33, 1996

[17] Weiskopf, N. G. & Chunhua, W. "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of Americam Medical Information Association,* Issue 20, pp. 144-151, 2013

[18] Wickramaarachchi,W.u & Karriaper.R.K. *An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis.* 2017 2nd International Conference on Image, Vision and Computing, IEEE, 2017.

[19] Yesha, Y., Janeja, V., Rishe, N. & Yesha, Y. "Personalized Decision Support System to Enhance Evidence Based Medicine through Big Data Analytics," *Healthcare Informatics (ICHI)*, 2014.

[20] Zolfaghar, K. et al. "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," *Big Data, 2013 IEEE International Conference on,* pp. 64-79, 2013.